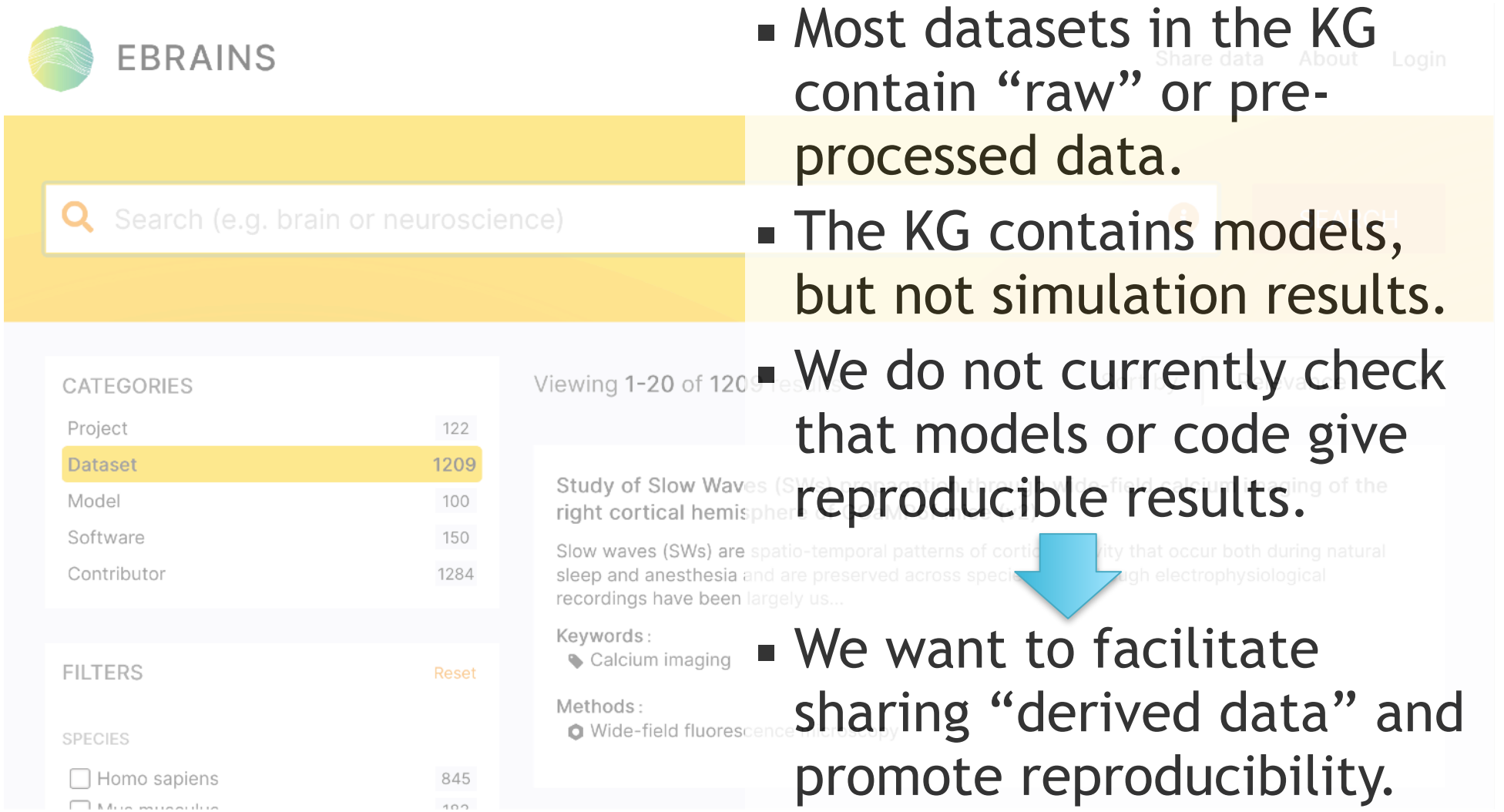
EBRAINS Provenance API

Andrew Davison
CNRS / Université Paris-Saclay

Reproducible research

- "I thought I used the same parameters but I'm getting different results"
- "I can't remember which version of the code I used to generate figure 6"
- "The new student wants to reuse that model I published three years ago but he can't reproduce the figures"
- "It worked yesterday"

Sharing simulations and data analysis results

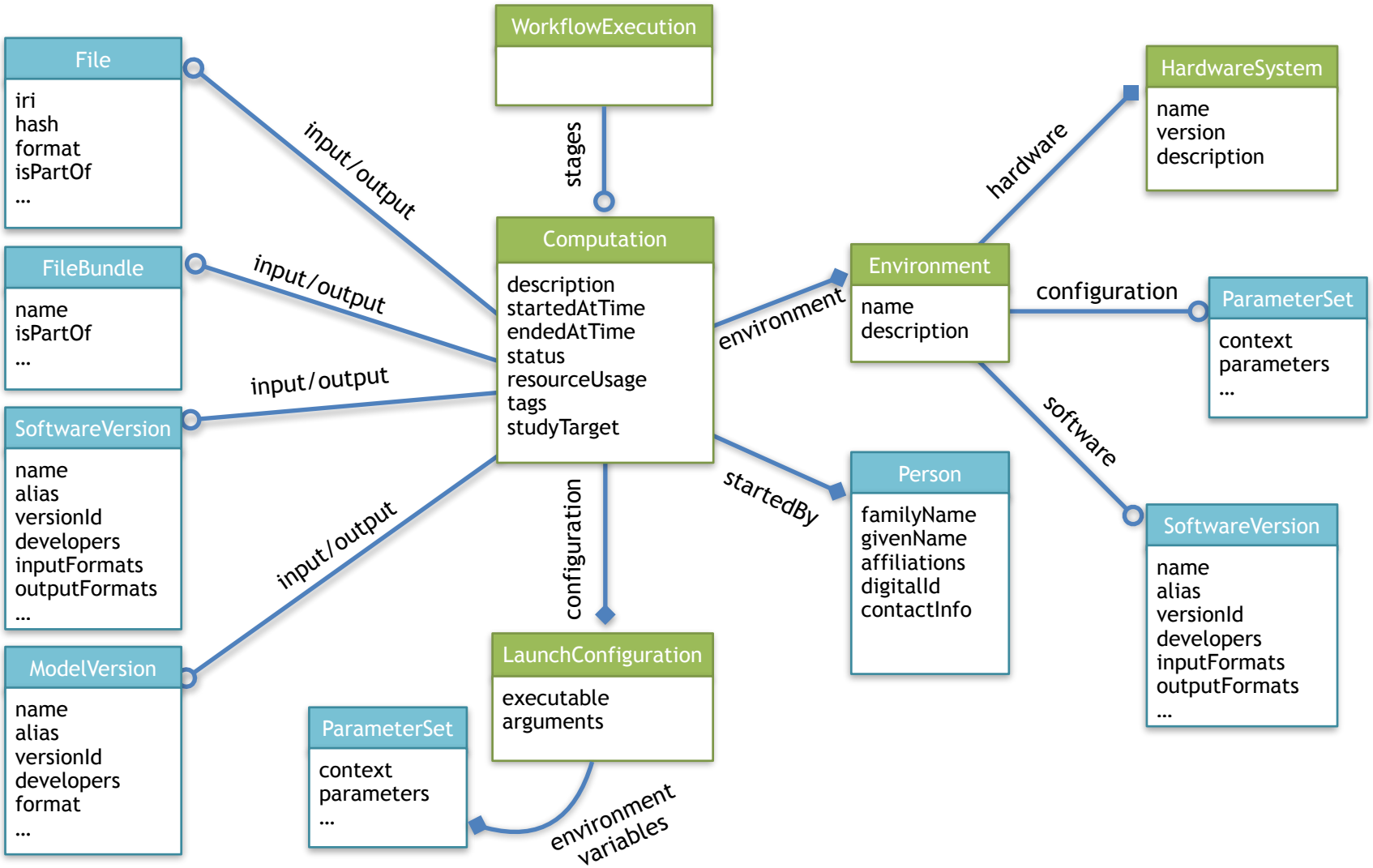


- Most datasets in the KG contain “raw” or pre-processed data.
- The KG contains models, but not simulation results.
- We do not currently check that models or code give reproducible results.
- We want to facilitate sharing “derived data” and promote reproducibility.

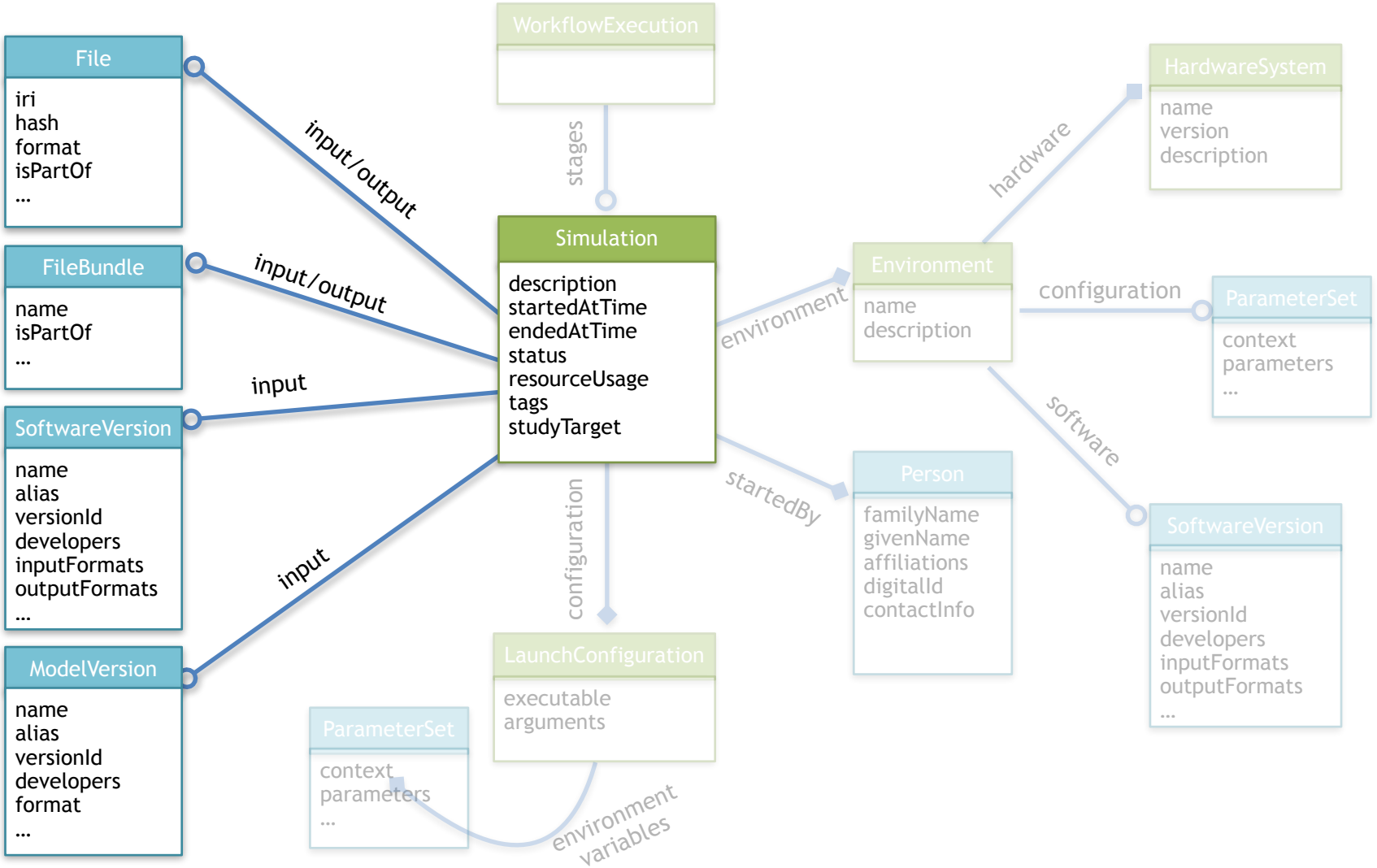
What information is needed?

- what code was run?
  - which executable?
    - name, location, version, compiler, compilation options
  - which script?
    - name, location, version
    - options, parameters
    - dependencies (name, location, version)
- what were the input data?
  - name, location, content
- what were the outputs?
  - data, logs, stdout/stderr
- who launched the computation?
- when was it launched/when did it run? (queueing systems)
- where did it run?
  - machine name(s), other identifiers (e.g. IP addresses)
  - processor architecture
  - available memory
  - operating system
- why was it run?
- what was the outcome?
- which project was it part of?

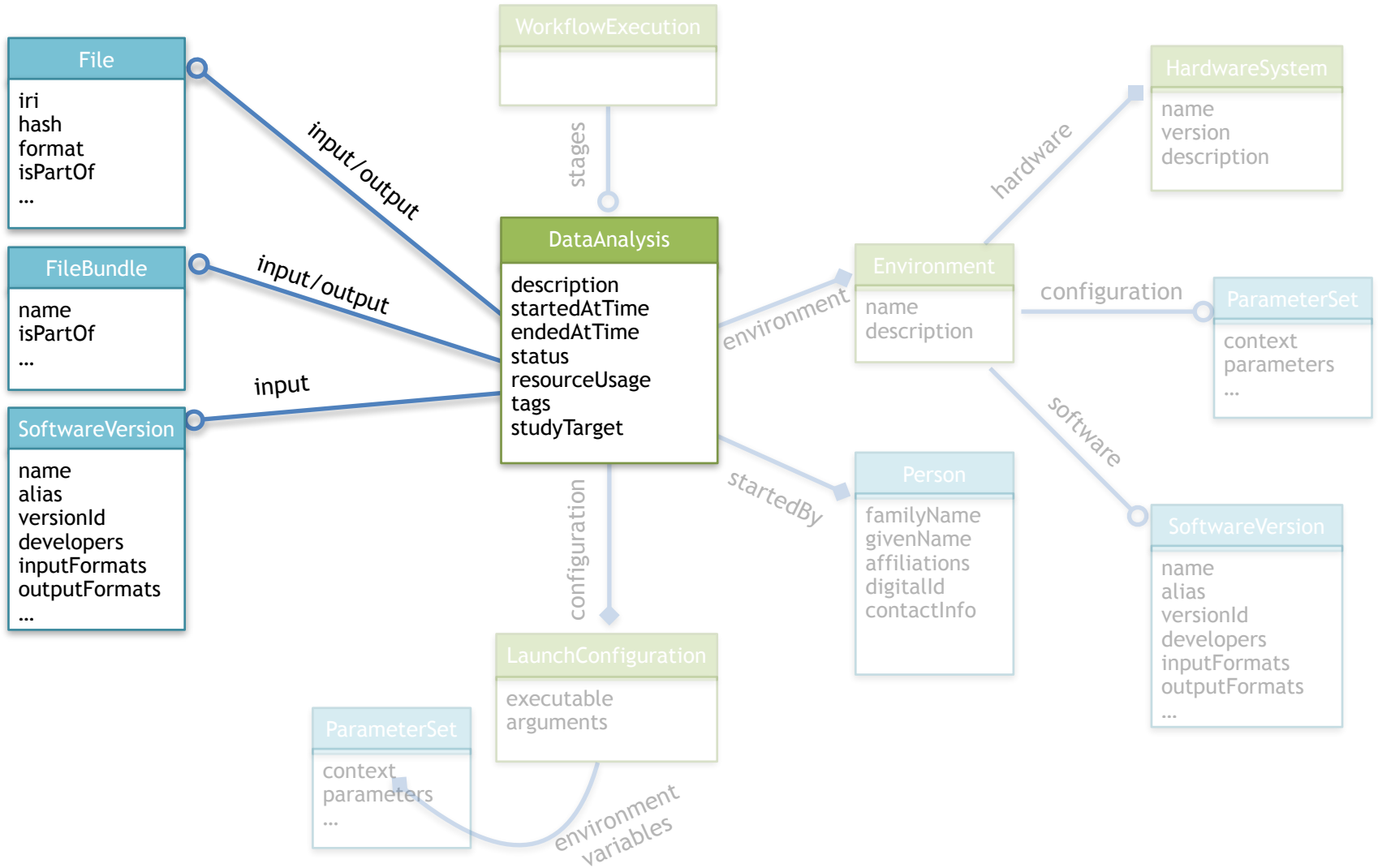
Knowledge Graph schemas



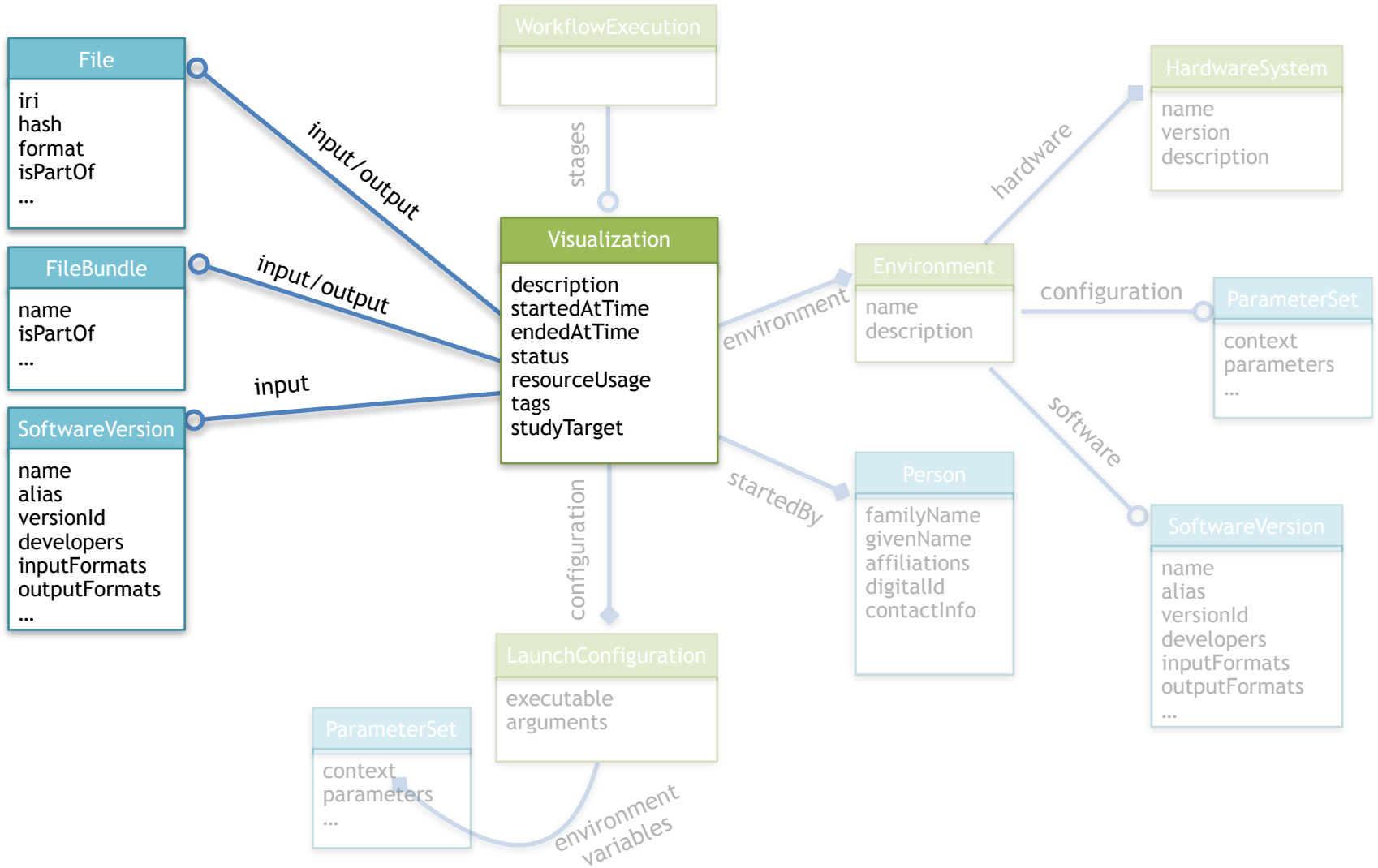
Knowledge Graph schemas



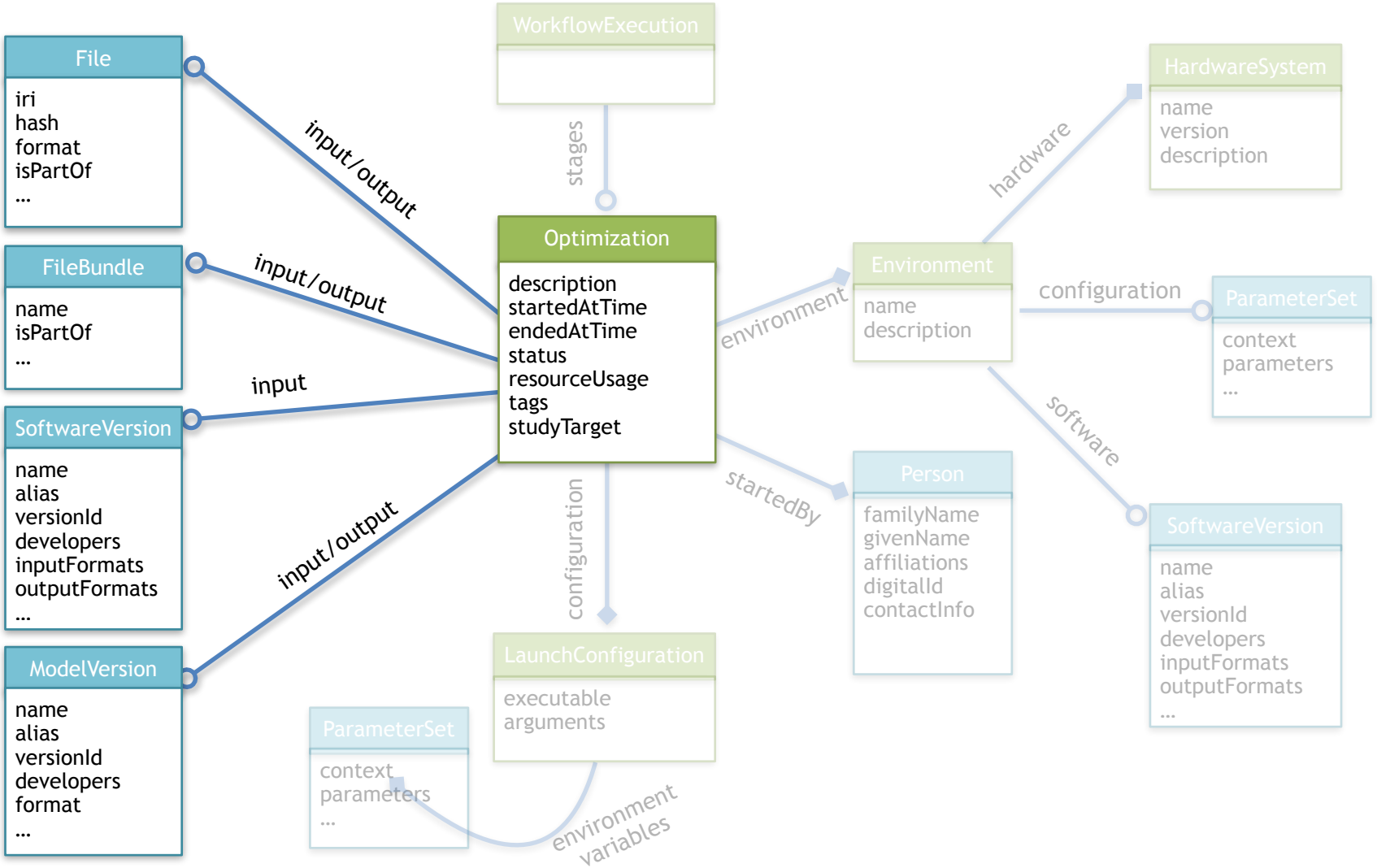
Knowledge Graph schemas



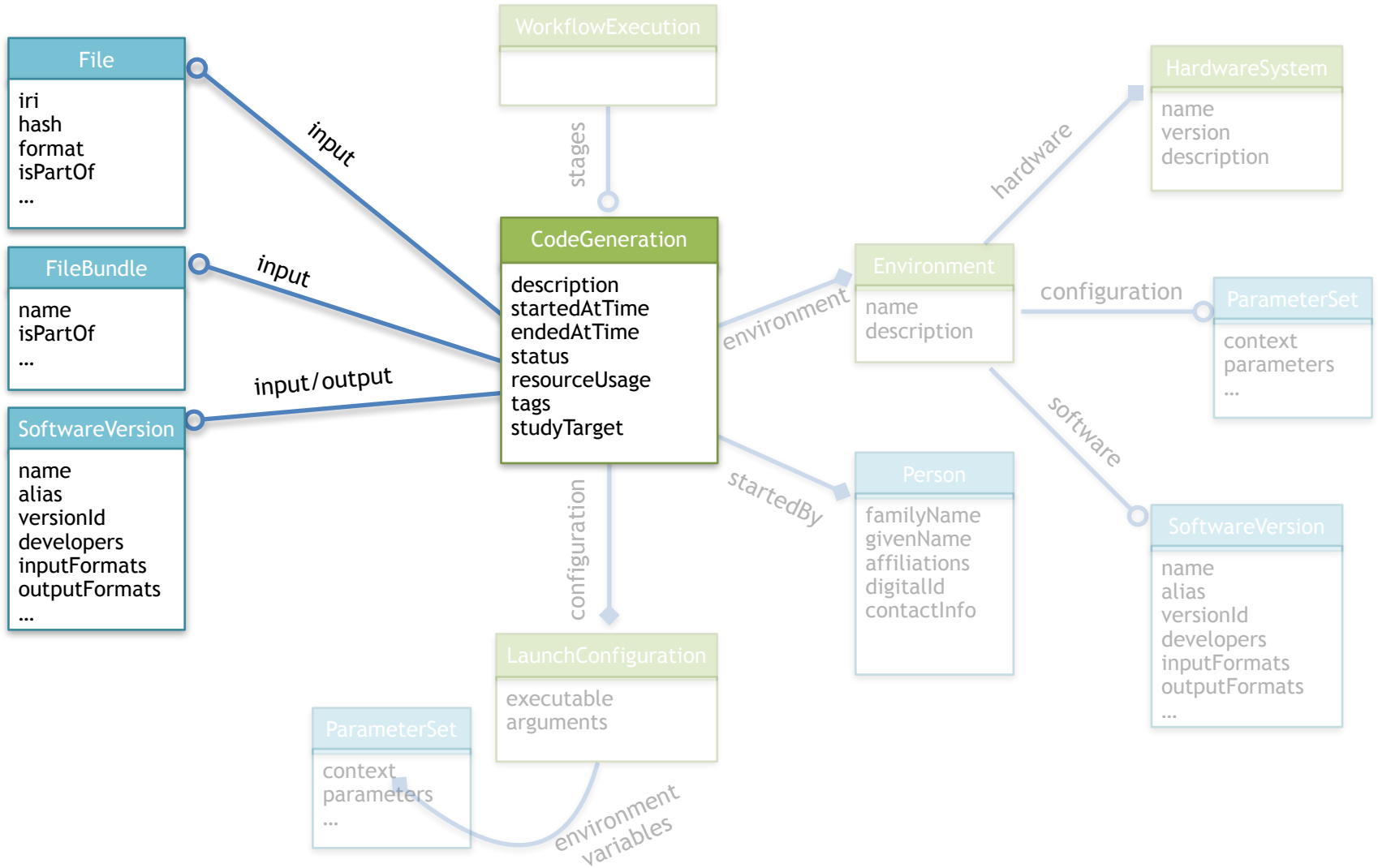
Knowledge Graph schemas



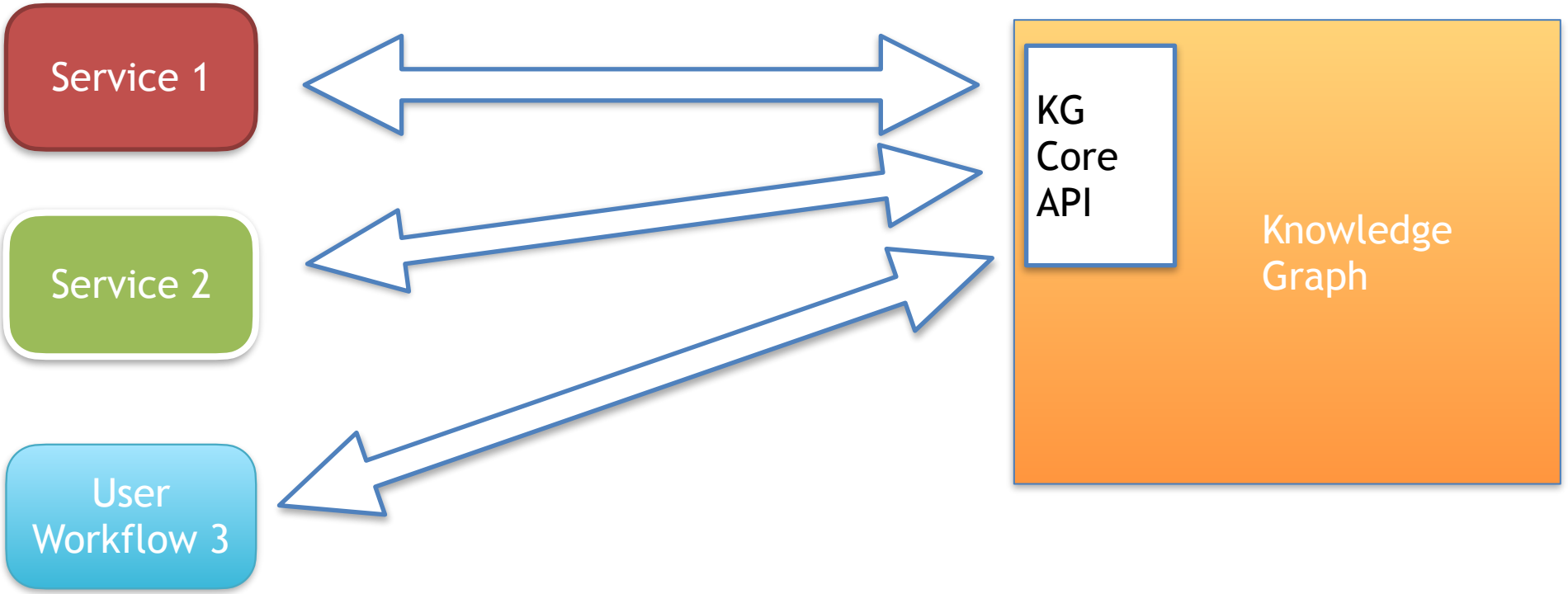
Knowledge Graph schemas



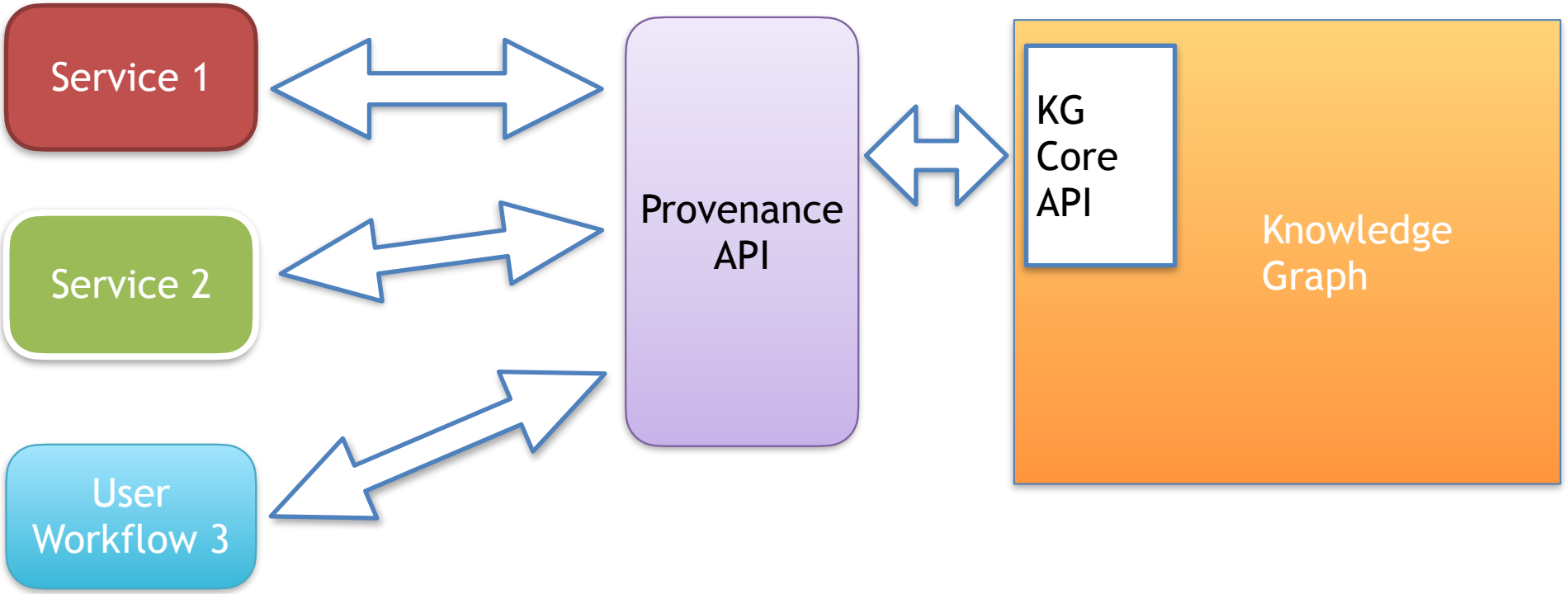
Knowledge Graph schemas



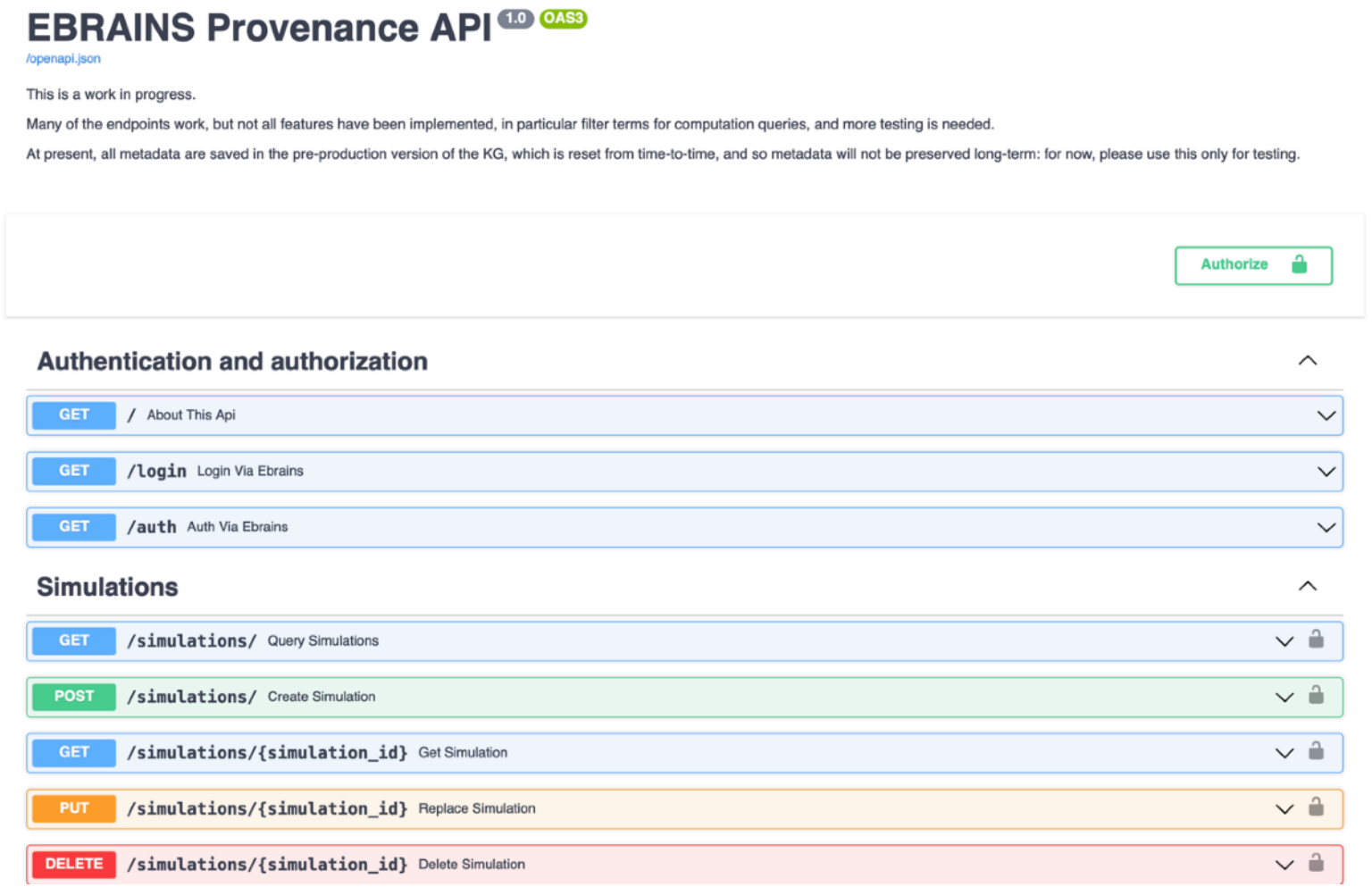
Communication between services/tools/workflows and the KG



Communication between services/tools/workflows and the KG



Provenance API - α release



https://prov.brainsimulation.eu/docs

Using the Provenance API

- “manual” use by individual users, e.g. in Jupyter notebooks
- integration into apps & services, e.g.
  - track simulations in Model Catalog app
  - migrate Neuromorphic job queue metadata to KG
  - capture all experiments run on the Neurorobotics Platform?
  - integrate into HPC job proxy?
- UNICORE integration / wrapper
- integration into CWL runners
- workflows dashboard
- provenance tracking of the Spack / Docker build process?!

Demo

https://wiki.ebrains.eu/bin/view/Collabs/ebrains-workflows

…/examples/visualisation/plot_ephys_data_with_prov_api.ipynb

Using / visualising provenance information